

Predicting Stroke

Stephen Kuc

2022-06-14

Importing dataset and necessary libraries

```
stroke <- read.csv("c:/Users/steph/OneDrive/Documents/USD/ADS503/healthcare-dataset-stroke-data.csv")
```

```
library(caret) # for training models
library(e1071)
library(Hmisc)
library(corrplot)
library(plyr)
library(pROC)
```

```
str(stroke)
```

```
## 'data.frame': 5110 obs. of 12 variables:
## $ id : int 9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
## $ gender : chr "Male" "Female" "Male" "Female" ...
## $ age : num 67 61 80 49 79 81 74 69 59 78 ...
## $ hypertension : int 0 0 0 0 1 0 1 0 0 0 ...
## $ heart_disease : int 1 0 1 0 0 0 1 0 0 0 ...
## $ ever_married : chr "Yes" "Yes" "Yes" "Yes" ...
## $ work_type : chr "Private" "Self-employed" "Private" "Private" ...
## $ Residence_type : chr "Urban" "Rural" "Rural" "Urban" ...
## $ avg_glucose_level: num 229 202 106 171 174 ...
## $ bmi : chr "36.6" "N/A" "32.5" "34.4" ...
## $ smoking_status : chr "formerly smoked" "never smoked" "never smoked" "smokes" ...
## $ stroke : int 1 1 1 1 1 1 1 1 1 1 ...
```

Many of the categorical variables are characters – we will need to change those to factors.

```
dim(stroke)
```

```
## [1] 5110 12
```

There are 5110 observations, with 12 features, including the target variable.

```
# changing datatypes to what they should be
stroke$hypertension <- as.factor(stroke$hypertension)
stroke$heart_disease <- as.factor(stroke$heart_disease)
```

```
stroke$gender <- as.factor(stroke$gender)
stroke$ever_married <- as.factor(stroke$ever_married)
stroke$work_type <- as.factor(stroke$work_type)
stroke$Residence_type <- as.factor(stroke$Residence_type)
stroke$smoking_status <- as.factor(stroke$smoking_status)
stroke$bmi <- as.numeric(stroke$bmi)
stroke$stroke <- as.factor(stroke$stroke)
```

```
# checking nulls
colSums(is.na(stroke))
```

```
##           id           gender           age           hypertension
##           0             0             0             0
## heart_disease ever_married work_type Residence_type
##           0             0             0             0
## avg_glucose_level      bmi smoking_status      stroke
##           0             201             0             0
```

There are 201 nulls in BMI.

```
summary(stroke)
```

```
##           id           gender           age           hypertension heart_disease
## Min.      : 67   Female:2994   Min.      : 0.08   0:4612       0:4834
## 1st Qu.:17741   Male  :2115   1st Qu.:25.00   1: 498       1: 276
## Median :36932   Other :    1   Median :45.00
## Mean    :36518                      Mean    :43.23
## 3rd Qu.:54682                      3rd Qu.:61.00
## Max.    :72940                      Max.    :82.00
##
## ever_married work_type Residence_type avg_glucose_level
## No :1757 children : 687 Rural:2514 Min. : 55.12
## Yes:3353 Govt_job : 657 Urban:2596 1st Qu.: 77.25
## Never_worked : 22 Median : 91.89
## Private :2925 Mean :106.15
## Self-employed: 819 3rd Qu.:114.09
## Max. :271.74
##
##           bmi           smoking_status stroke
## Min.      :10.30 formerly smoked: 885 0:4861
## 1st Qu.:23.50 never smoked :1892 1: 249
## Median :28.10 smokes : 789
## Mean :28.89 Unknown :1544
## 3rd Qu.:33.10
## Max. :97.60
## NA's :201
```

1 “other” gender. 1544 “unknown” smoker status. 201 nulls in BMI. Work type “Private” means what? Any cutoff for minimum Age? Target variable Stroke seems pretty imbalanced.

```

# plots for all features
par(mar = c(2,2,2,2))
layout.matrix <- matrix(c(1,4,5,6,2,7,8,9,3,10,0,0),nrow = 4, ncol = 3)

layout(mat = layout.matrix,
       heights = c(4, 4, 4, 4),
       widths = c(3, 3, 3))
# histogram for numerical features
hist(stroke$age, cex.main = .5, cex.axis = .5)
hist(stroke$avg_glucose_level, cex.main = .5)
hist(stroke$bmi, cex.main = .5)

# bar charts for categorical
countGen <- table(stroke$stroke,stroke$gender)
barplot(countGen, main = "Stroke distribution by Gender", legend = rownames(countGen), cex.lab = .5, cex.axis = .5)

countHyp <- table(stroke$stroke, stroke$hypertension)
barplot(countHyp, main = "Stroke distribution by Hypertension", legend = rownames(countHyp), cex.lab = .5, cex.axis = .5)

countHd <- table(stroke$stroke,stroke$heart_disease)
barplot(countHd, main = "Stroke distribution by Heart Disease", legend = rownames(countHd), cex.lab = .5, cex.axis = .5)

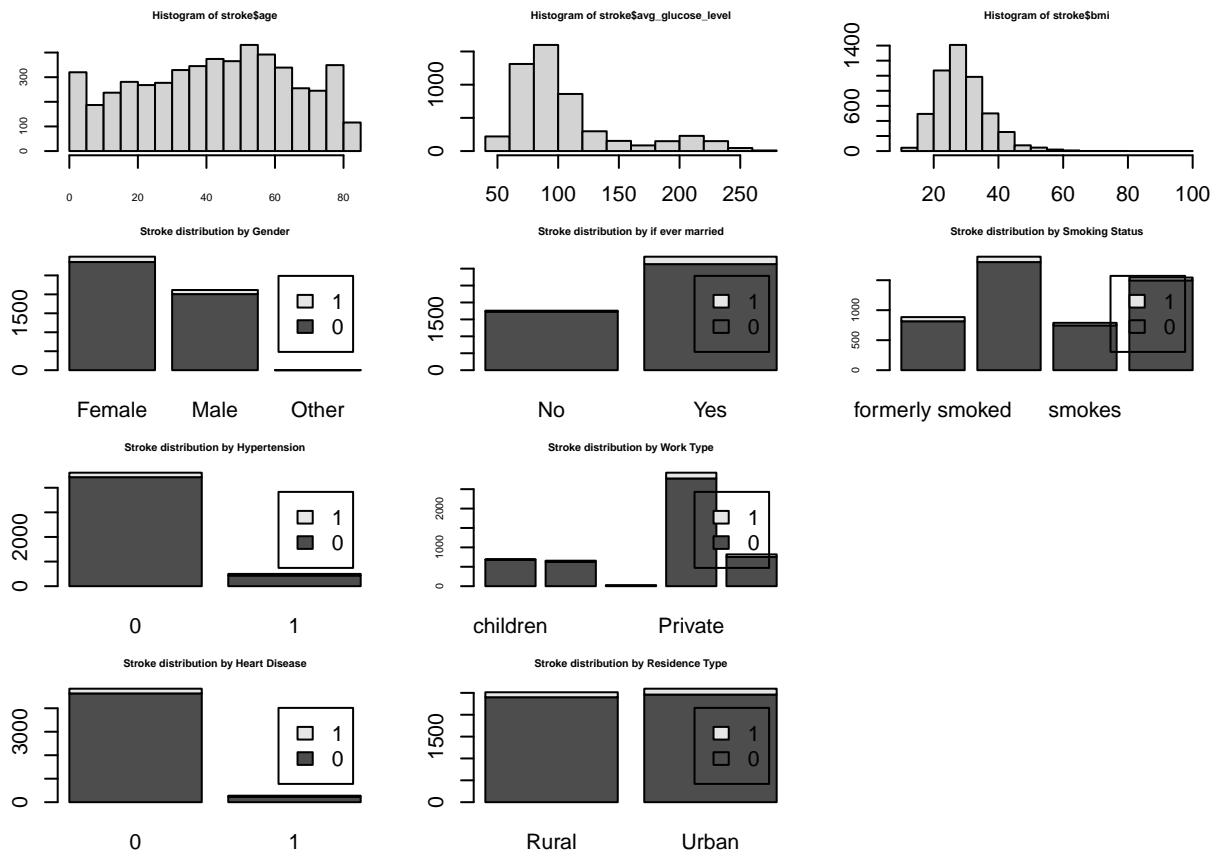
countMarried <- table(stroke$stroke, stroke$ever_married)
barplot(countMarried, main = "Stroke distribution by if ever married", legend = rownames(countMarried), cex.lab = .5, cex.axis = .5)

countWork <- table(stroke$stroke, stroke$work_type)
barplot(countWork, main = "Stroke distribution by Work Type", legend = rownames(countWork), cex.lab = .5, cex.axis = .5)

countRes <- table(stroke$stroke, stroke$Residence_type)
barplot(countRes, main = "Stroke distribution by Residence Type", legend = rownames(countRes), cex.lab = .5, cex.axis = .5)

countSmoke <- table(stroke$stroke, stroke$smoking_status)
barplot(countSmoke, main = "Stroke distribution by Smoking Status", legend = rownames(countSmoke), cex.lab = .5, cex.axis = .5)

```



```
# glucose levels look skewed slightly, as does bmi numbers
# let's check for skewness
skewness(stroke$avg_glucose_level)
```

```
## [1] 1.571361
```

```
skewness(stroke$bmi) ## need to get rid of nulls to see skewness metric
```

```
## [1] NA
```

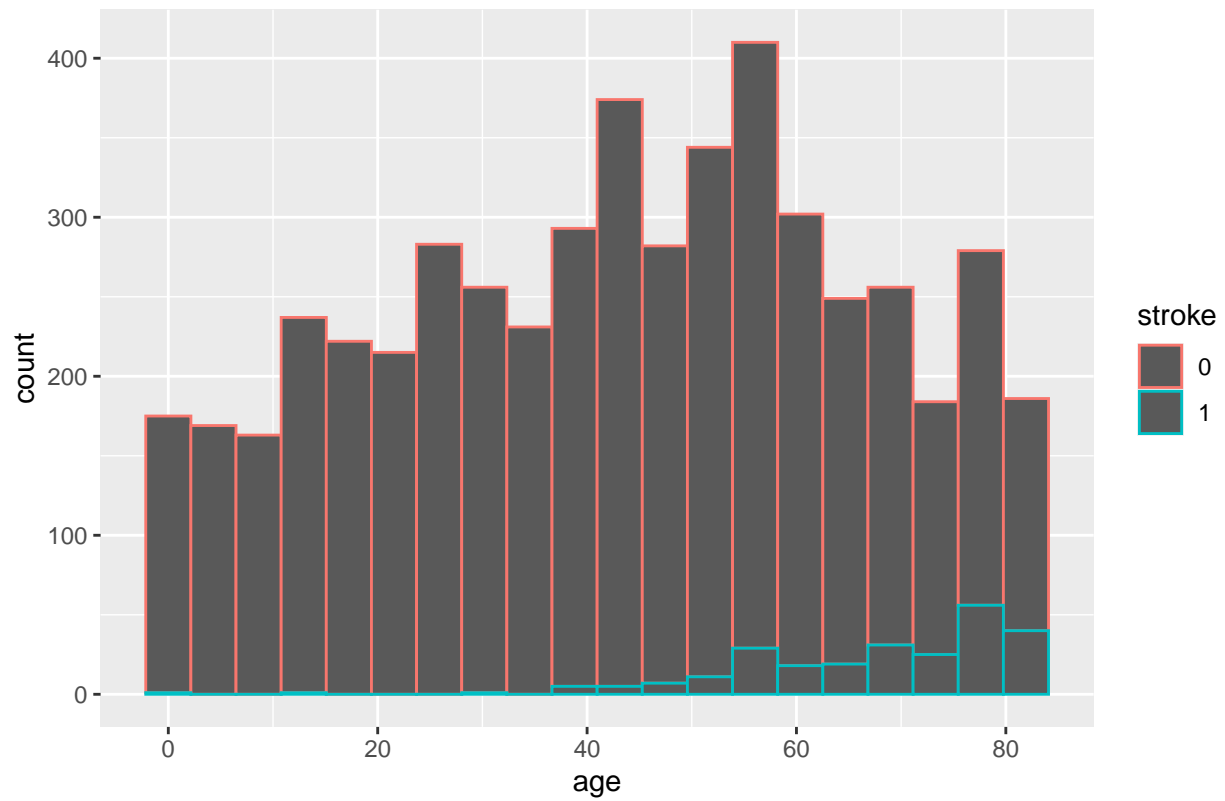
```
skewness(stroke$age)
```

```
## [1] -0.1369789
```

Slight skew in glucose level.

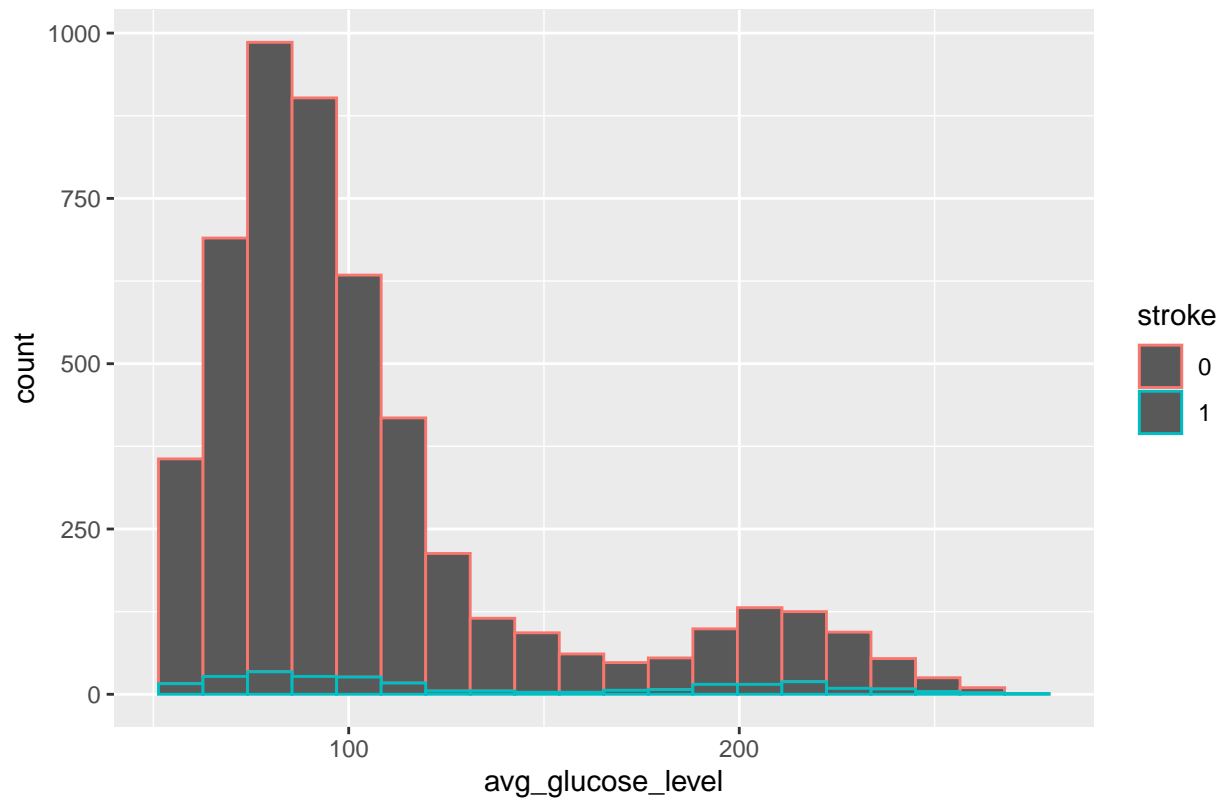
```
# let's investigate the numeric variables further
ggplot(stroke, aes(x=age, color=stroke)) + geom_histogram(bins = 20) + ggtitle("Histogram of age and stroke")
```

Histogram of age and stroke occurrence



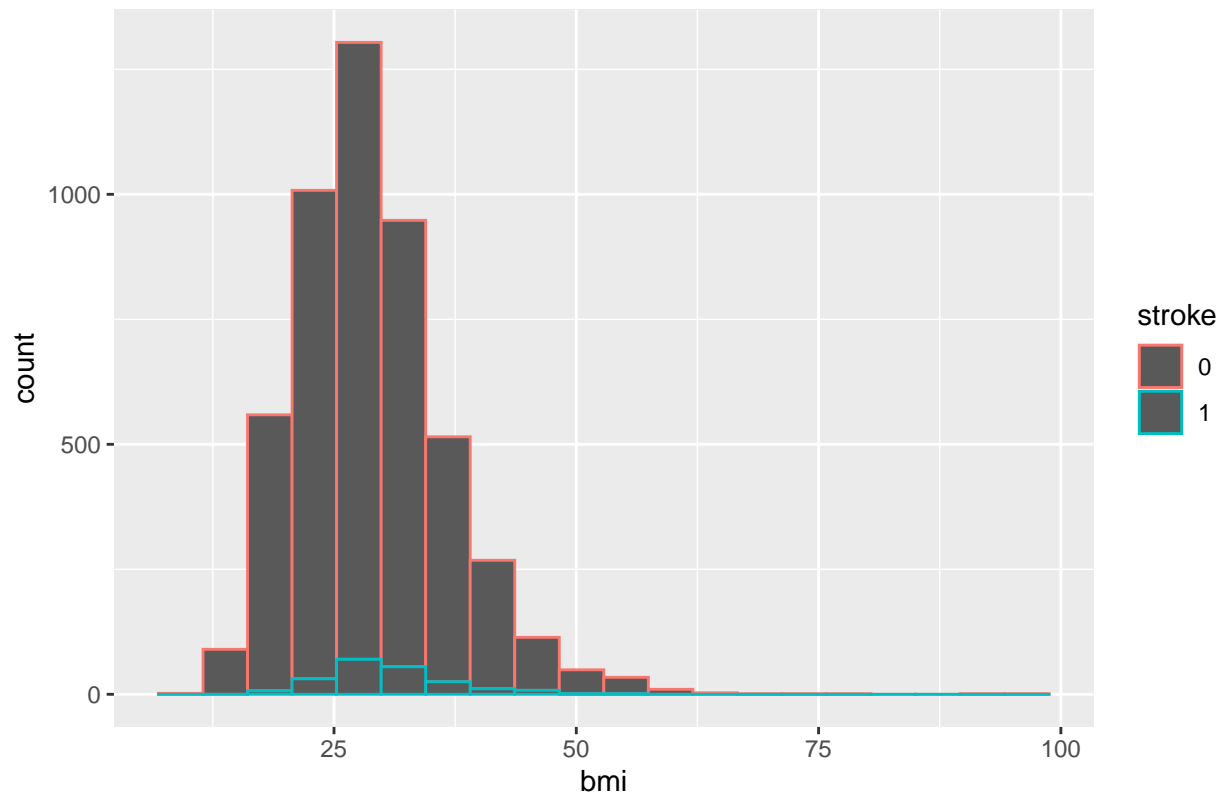
```
ggplot(stroke, aes(x=avg_glucose_level,color=stroke)) + geom_histogram(bins = 20) + ggtitle("Histogram of age and stroke occurrence")
```

Histogram of average glucose level and stroke occurrence



```
ggplot(stroke, aes(x=bmi,color=stroke)) + geom_histogram(bins = 20) + ggtitle("Histogram of bmi and stroke")
```

Histogram of bmi and stroke occurrence



creating a contingency table of categorical variables

```
prop.table(table(stroke$ever_married, stroke$stroke),margin=1)*100
```

```
##
##           0           1
## No  98.349459  1.650541
## Yes  93.438712  6.561288
```

```
prop.table(table(stroke$gender, stroke$stroke),margin=1)*100
```

```
##
##           0           1
## Female  95.290581  4.709419
## Male    94.893617  5.106383
## Other   100.000000  0.000000
```

```
prop.table(table(stroke$work_type, stroke$stroke),margin=1)*100
```

```
##
##           0           1
## children  99.7088792  0.2911208
## Govt_job  94.9771689  5.0228311
## Never_worked 100.0000000  0.0000000
## Private   94.9059829  5.0940171
## Self-employed 92.0634921  7.9365079
```

```
prop.table(table(stroke$heart_disease, stroke$stroke),margin=1)*100
```

```
##
##           0           1
##  0 95.821266  4.178734
##  1 82.971014 17.028986
```

```
prop.table(table(stroke$Residence_type, stroke$stroke),margin=1)*100
```

```
##
##           0           1
##  Rural 95.465394  4.534606
##  Urban 94.799692  5.200308
```

```
prop.table(table(stroke$hypertension, stroke$stroke),margin=1)*100
```

```
##
##           0           1
##  0 96.03209  3.96791
##  1 86.74699 13.25301
```

```
prop.table(table(stroke$smoking_status, stroke$stroke),margin=1)*100
```

```
##
##           0           1
##  formerly smoked 92.090395  7.909605
##  never smoked   95.243129  4.756871
##  smokes          94.676806  5.323194
##  Unknown        96.955959  3.044041
```

```
# convert factors into dummy var
```